J. Clin. Chem. Clin. Biochem.
Vol. 18, 1980, pp. 621–625

The Use of Patient Data for the Calculation of Reference Values for Some Haematological Parameters

By *A. J. Naus*, *A. Borst* and *P. S. Kuppens*

Department of Hematology and Clinical Chemistry, St. Laurentius Hospital, Mgr. Driessenstraat 6, 6043 CV Roermond, The Netherlands

(Received January 28/June 18, 1980)

Summary: We have investigated the use of patient data for the calculation of reference values for the parameters which are determined by the Hemalog. For this purpose we used the *Bhattacharya* plot. All the parameters, with the exception of leukocytes, appear to meet the main underlying assumption of this plot, namely that the fequency distribution is *Gauss*ian. In the case of leukocytes, however, the frequency distribution could be resolved into two overlapping *Gauss*ian curves, thus making it possible to calculate reference values for this parameter also.

The reference values as calculated from 14,500 unselected data (excluding children) are in general agreement with the literature. Significant differences were however detected between a group of patients and a group of blood donors.

When a *Bhattacharya* plot has to be constructed with relatively few data, smoothing of the observed frequencies is very helpful in deciding which part of the plot is linear. Smoothing was carried out using the least squares method with a quadratic equation. Since the classes are equally spaced, this involves only a simple numerical transformation of the frequencies.

Die Verwendung von Patienten-Daten für die Ermittlung von Referenzwerten für einige hämatologische Kenngrößen

Zusammenfassung: Wir haben die Verwendung von Patienten-Daten für die Ermittlung von Referenzwerten für Kenngrößen, die mit dem Hemalog bestimmt wurden, untersucht und dafür die *Bhattacharya*-Darstellung benutzt. Alle Kenngrößen außer den Leukocyten scheinen die dieser Darstellung hauptsächlich zugrundeliegende Annahme, daß eine *Gauss*'-Verteilung vorliegt, zu erfüllen. Für die Leukocyten konnte die Häufigkeitsverteilung in zwei sich überlappende *Gauss*-Kurven aufgelöst werden, so daß auch für diese Kenngröße Referenzwerte ermittelt werden konnten.

Die aus 14.500 unausgewählten Daten (Kinder ausgenommen) ermittelten Referenzwerte stimmen mit Literaturangaben überein. Signifikante Unterschiede wurden jedoch zwischen einer Gruppe von Patienten und einer Gruppe von Blutspendern gefunden.

Wenn die *Bhattacharya*-Darstellung aus relativ wenig Daten konstruiert wird, ist die Glättung der beobachteten Häufigkeiten für die Entscheidung, welcher Teil der Darstellung linear ist, sehr hilfreich. Die Glättung wurde unter Verwendung der Methode der kleinsten Quadrate mit einer quadratischen Gleichung durchgeführt. Da die Klassen gleichen Raum einnehmen, beinhaltet dies nur eine einfache numerische Transformation der Häufigkeiten.

Introduction

The Hemalog (Technicon, Tarrytown, New York), determines simultaneously in a blood sample platelets (PLTS), leukocytes (WBC), erythrocytes (RBC), haemoglobin (Hb) and packed cell volume (PCV) and calculates the mean corpuscular volume MCV = PCV/RBC), mean corpuscular haemoglobin (MCH = Hb/RBC) and mean corpuscular haemoglobin concentration (MCHC = Hb/PCV = MCH/MCV) (1).

During the development of a quality control program for the Hemalog, the results of which will be reported in a subsequent paper, the need was felt for reference values of greater accuracy than those currently available. The problems, however, encountered in finding a group of persons that can be used for the determination of these values are numerous (2). It certainly is not acceptable in our view to use the laboratory staff or a group of blood donors for this purpose, because they do not form a true representation of the whole population. In case of

0340-076X/80/0018-0621$2.00

the laboratory staff for instance, the majority of people are female and below the age of 30. Most of our blood donors are male and between 30 and 40. So automatically a selection is made when choosing one of these groups for the calculation of reference ranges. The danger that these ranges are biased when a selection is made beforehand is very great. So in our view it is better to make no selection at all. Simply take all the results produced during a certain time in your laboratory for a certain test and use these. Of course a number of these results are "abnormal" and should not be used for the calculation of mean and standard deviation.

Applying the *Bhattacharya* plot automatically means that this is achieved (3). The only assumption that has to be made when using this plot is that the frequency distribution is *Gauss*ian. If not, the plot cannot be applied.

However, the number of abnormal results in the population used for the calculation of reference ranges can be very great. In these cases the method of *Hoffmann* (4) gives an S-shaped curve as was discussed by *White* (5).

Because no selection is made a true representation of the population can be expected. This is even more so, because a very large number of test results (of which the majority are "normal") can be accumulated in a relatively short time.

Materials and Methods

The essence of the *Bhattacharya* plot is the following: The results for a certain assay are accumulated in classes, which must be equally spaced. The logarithm of the quotient of the frequencies in class i + 1 and class i is then plotted against the midpoint of class i, partly resulting in a straight line. This straight line represents the part of the distribution that is truly *Gauss*ian, excluding all "abnormal" results. The slope of this line and the intercept with the X-axis result in the standard deviation and the mean of the distribution respectively. Mathematically it can be expressed as follows.

$$f_i = \frac{N}{\sigma\sqrt{2\,n}} \exp\left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right), \qquad \text{(eq 1)}$$

where:

x_i = the midpoint of class i.
f_i = the number of results in class i.
N = the total number of results in the *Gauss*ian distribution.
μ = the mean of the distribution
σ = the sd of the distribution.

Equation 1 is the well known frequency density function for a *Gauss*ian distribution from which the following can be easily derived.

$$\log\frac{f_{i+1}}{f_i} = -\frac{1}{2}\frac{h^2}{\sigma^2} - \frac{h(x_i-\mu)}{\sigma^2}, \qquad \text{(eq 2)}$$

where:

h = the width of the classes.

Equation 2 represents a straight line with $x = x_i$ and $y = \log\frac{f_{i+1}}{f_i}$.

The intercept with the X-axis is found for $x = \mu - 1/2\,h$. The slope of this line equals $-h/\sigma^2$, hence $\sigma^2 = -h/\text{slope}$. However, due to the grouping of data a correction must be applied. This results in.

$$\sigma^2 = -h/\text{slope} - h^2/12, \qquad \text{(eq 3)}$$

where:

$-h^2/12$ is *Sheppards* correction.

The *Bhattacharya* plot is only linear for the part of the frequency distribution that is truly *Gauss*ian. At the high and low end the deviations from linearity due to "abnormal" results are in most cases very obvious. To calculate reference ranges using this plot, the linear part is selected manually and through these points a straight line is calculated using the method of least squares. When the number of data is small, the selection of the straight part in the *Bhattacharya* plot can sometimes be difficult due to statistical variations. In these cases a smoothing procedure is indicated. This facilitates the selection of the points of the plot that lie on the straight part and ensures that "abnormal" results (giving rise to the non-linear part of the plot) are not taken into account.

The smoothing procedure must be carried out on the observed frequencies and not on the points of the *Bhattacharya* plot itself, because the y-values of this plot are not independent. Suppose for instance that a point in the *Bhattacharya* plot has a y_i value that is too high, as a result of statistical variations. Since $y_i = \log\frac{f_{i+1}}{f_i}$ this means that f_{i+1} is too high, f_i too low, or both. It further means that correcting y_i automatically changes y_{i+1}, y_{i-1} or both, since the denominator of y_i is the nominator in y_{i-1} and the nominator of y_i is the denominator in y_{i+1}.

Smoothing of the frequency values was done according to *Savitzky* et al. (6, 7). Five consecutive frequencies (f_{i-2}, f_{i-1}, f_i, f_{i+1}, f_{i+2}) are fitted to a parabola using the method of least squares. The frequency read from the parabola at x_i is the smoothed frequency. Next a parabola is calculated through the frequencies at x_{i-1}, x_i, x_{i+1}, x_{i+2} and x_{i+3} and the value read from the parabola at the midpoint (x_{i+1}) again is the smoothed frequency and so on. Since the classes are equally spaced, the smoothed frequencies can simply be calculated using the following expression.

$$f_{i,\,\text{smoothed}} = \frac{-3f_{i-2} + 12f_{i-1} + 17f_i + 12f_{i+1} - 3f_{i+2}}{35} \qquad \text{(eq 4)}$$

An example of the difference between a smoothed and unsmoothed *Bhattacharya* plot is given in figure 1. The corresponding data are summarized in table 1.

Since the frequency distribution for leukocytes appeared to be very skewed, the possibility was investigated that this distribution could be described by the sum of two overlapping *Gauss*ian curves.

The procedure that has to be used when overlap occurs is as follows. First, using the observed frequencies, a *Bhattacharya* plot is constructed. The linear part of this plot results in values for the mean and standard deviation of a *Gauss*ian distribution. The calculated frequencies of this distribution are now subtracted from the observed ones. With the resulting differences a new *Bhattacharya* plot is constructed. If a second *Gauss*ian distribution is hidden under the first one, a straight part in this plot can be seen. From this straight part the mean and standard deviation of a second *Gauss*ian distribution can be calculated.

When this procedure is carried out for leukocytes, it appears that the frequency distribution can indeed be described by the sum of two overlapping *Gauss*ian curves ($\mu_1 = 6.55$, $sd_1 = 1.56$, $\mu_2 = 9.86$, $sd_2 = 1.75$). The details of the calculation are given in table 2 and the resulting frequency distributions are depicted in figure 2.

Tab. 1. Data for figure 1.
Frequency distribution for 1593 mean corpuscular volume (MCV) values. The unsmoothed and smoothed *Bhattacharya* plot are shown in figure 1.
The calculation of the smoothed line gives the following results (figures indicated by an asterisk) r = - 0.9997, slope = - 0.1018 and intercept with X axis = 91.847 and from this μ = 92.85 and sd = 4.40. Units of X:fl/cell.

X	F	dLN	Fs-moothed	dLN
81	13	0.932	–	–
83	33	0.288	–	–
85	44	1.012	57	0.702*
87	121	0.435	115	0.481*
89	187	0.245	186	0.288*
91	239	0.190	248	0.100*
93	289	- 0.182	274	- 0.116*
95	241	- 0.337	244	- 0.327*
97	172	- 0.368	176	- 0.408
99	119	- 0.531	117	- 0.486
101	70	- 0.585	72	- 0.693
103	39	- 1.099	36	–
105	13	0.000	–	–
107	13	–	–	–

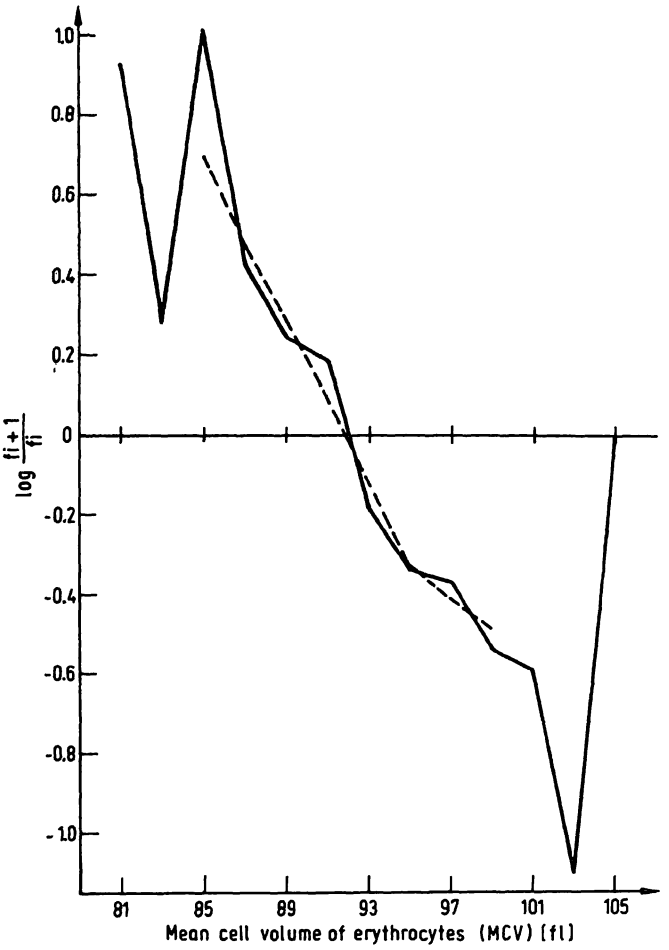

Fig. 1. *Bhattacharya* plot for mean corpuscular volume (MCV) with (dotted line) and without (full line) smoothing.

When two populations are compared, μ_1 is said to be statistically different from μ_2 when the 95% confidence limits of μ_1 and μ_2 do not overlap. For the standard deviations of two populations an analogous statement can be made.

The 95% confidence limits for the standard deviation can be calculated from the variance of the slope of the linear part of the *Bhattacharya* plot.

$$\sqrt{\left(\frac{-h}{\text{slope} - 2\text{sd}_{\text{slope}}} - \frac{h^2}{12}\right)} < \text{sd} < \sqrt{\left(\frac{-h}{\text{slope} + 2\text{sd}_{\text{slope}}} - \frac{h^2}{12}\right)} \qquad \text{(eq 5)}$$

The variance of the slope was calculated using the following equation (10).

$$\text{sd}^2_{\text{slope}} = \frac{\text{sd}_y^2}{\text{sd}_x^2} \frac{(1 - r^2)}{(n - 2)}, \qquad \text{(eq 6)}$$

Tab. 2. Resolution of the frequency distribution of leukocytes (WBC) into two *Gaussian* distributions.
F (column 2) are the observed frequencies. F_{corr} are the observed frequencies smoothed according to eq. 4. $F_{calc\ 1}$ are the calculated frequencies of the first *Gaussian* curve (μ_1 = 6.55, sd_1 = 1.56). Differences are the differences between the frequencies in columns 2 and 5 respectively. $F_{calc\ 2}$ are the calculated frequencies of the second *Gaussian* curve (μ_2 = 9.86, sd_2 = 1.75). $F_{tot\ calc}$ are the sums of the frequencies in columns 5 and 8.
The points indicated by an asterisk (columns 4 and 7) have been taken for the calculation of the straight line.
For further explanation see the text.

Interval mean	F	$F_{corr.}$	dLN	$F_{calc\ 1}$	Difference	dLN	$F_{calc\ 2}$	$F_{tot\ calc}$
1.5	53	–	–	13.9			0.0	13.9
2.5	135	–	–	87.5			0.1	87.6
3.5	343	393.6	1.027*	375.8			1.3	377.1
4.5	1053	1099.1	0.603*	1071.7			9.1	1080.8
5.5	2051	2009.4	0.229*	2027.2			44.7	2071.9
6.5	2585	2525.3	- 0.051	2541.2			157.5	2698.7
7.5	2388	2398.5	- 0.230	2111.7	286.8	0.948	400.7	2512.4
8.5	1889	1904.1	- 0.292	1164.2	739.9	0.296*	735.6	1899.8
9.5	1395	1420.6	- 0.342	425.5	995.1	- 0.095*	974.0	1399.5
10.5	1037	1008.2	- 0.421	103.4	904.8	- 0.339*	930.5	1033.9
11.5	647	661.4	- 0.502	16.9	644.5	- 0.476	641.3	658.2
12.5	393	400.4	–	0.0	400.4	–	318.8	318.8
13.5	275	–	–	0.0	–	–	114.4	114.4
14.5	211	–	–	0.0	–	–	29.6	29.6

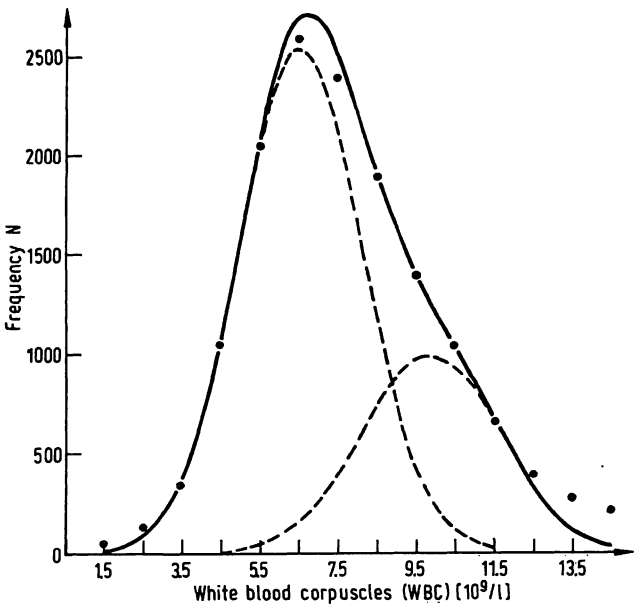

Fig. 2. Resolution of the frequency distribution of leukocytes (WBC) into two overlapping *Gaussian* curves (dotted lines). The full line represents the sum of the frequencies. • are the observed frequencies.

where:

y = interval mean.

$x = \log \frac{f_{i+1}}{f_i}$.

r = coefficient of correlation.

n = number of points on the linear part.

The 95% confidence limits for the mean have been calculated as follows.

$$\text{intcpt} + \frac{1}{2}h - 2sd_{\text{intcpt}} < \mu < \text{intcpt} + \frac{1}{2}h + 2sd_{\text{intcpt}} \quad \text{(eq 7)}$$

The variance of the intercept can be calculated with the following expression.

$$sd^2_{\text{intcpt}} = sd^2_{\text{slope}}\left(sd^2_x \frac{(n-1)}{n} + \mu^2_x\right) \quad \text{(eq 8)}$$

For the collection of data and all calculations, use was made of a Wang PCS II with punched tape reader.

Results and Discussion

By comparing the reference values calculated from a group of hospitalized patients (N = 3500) and a group of blood donors (N = 650), the following conclusions can be drawn.

1. The mean is the same for the two groups for all parameters. For female blood donors, however, the mean is slightly but significantly higher than for patients (tab. 3).

2. The standard deviation for all parameters is smaller for blood donors than for patients (tab. 4).

Because the *Bhattacharya* plot selects in fact the haematologically "normal" people, it is not surprising in

Tab. 3. Comparison between patients and blood donors – mean values. For explanation see the text.

		Patients		Blood donors	
		μ	Range μ	μ	Range μ
PLTS	(10^9/l)	167.5	166.2 – 168.8	169.3	168.7 – 169.8
RBC ♂	(10^{12}/l)	5.07	5.05 – 5.09	5.09	5.07 – 5.11
RBC ♀	(10^{12}/l)	4.56	4.51 – 4.61	4.64	4.62 – 4.66
Hb ♂	(Fe, mmol/l)	9.93	9.85 – 10.01	9.90	9.86 – 9.90
Hb ♀	(Fe, mmol/l)	8.81	8.79 – 8.83	8.91	8.86 – 8.96
PCV ♂	(l/l)	0.471	0.4686 – 0.4726	0.472	0.4704 – 0.4730
PCV ♀	(l/l)	0.432	0.4307 – 0.4333	0.436	0.4334 – 0.4380
MCV	(fl)	93.5	93.19 – 93.77	93.6	93.51 – 93.73
MCH	(Fe, fmol)	1.94	1.940 – 1.942	1.94	1.938 – 1.938
MCHC	(Fe, mmol/l)	21.0	20.92 – 21.00	21.0	20.91 – 20.99

Tab. 4. Comparison between patients and blood donors – standard deviations. For explanation see the text.

		Patients		Blood donors	
		SD	Range SD	SD	Range SD
PLTS	(10^9/l)	43.7	42.7 – 44.8	42.3	42.0 – 42.6
RBC ♂	(10^{12}/l)	0.48	0.46 – 0.51	0.39	0.38 – 0.40
RBC ♀	(10^{12}/l)	0.44	0.42 – 0.47	0.35	0.35 – 0.36
Hb ♂	(Fe, mmol/l)	0.66	0.62 – 0.72	0.48	0.45 – 0.52
Hb ♀	(Fe, mmol/l)	0.70	0.69 – 0.71	0.55	0.53 – 0.57
PCV ♂	(l/l)	0.0302	0.0281 – 0.0329	0.0229	0.0223 – 0.0235
PCV ♀	(l/l)	0.0311	0.0300 – 0.0323	0.0269	0.0257 – 0.0284
MCV	(fl)	4.77	4.56 – 5.01	4.11	4.02 – 4.19
MCH	(Fe, fmol)	0.099	0.098 – 0.101	0.091	0.090 – 0.091
MCHC	(Fe, mmol/l)	0.70	0.67 – 0.74	0.58	0.54 – 0.62

Tab. 5. Reference values.
The reference values have been calculated as $\mu \pm 2SD$,
the reference range for WBC has been calculated as $\mu_1 - 1.8\ SD_1$ $\mu_2 + 1.4\ SD_2$.

		Our method	Ref. (8)	Ref. (9)
PLTS	(10^9/l)	130 –320	140 –440	140 –340
WBC	(10^9/l)	3.8 – 12.3	4.3 – 10.0	4.0 – 10.0
RBC ♂	(10^{12}/l)	4.1 – 6.0	4.5 – 6.3	4.5 – 6.5
RBC ♀	(10^{12}/l)	3.7 – 5.4	4.2 – 5.5	3.9 – 5.6
Hb ♂	(Fe, mmol/l)	8.5 – 11.2	8.7 – 11.2	8.4 – 11.2
Hb ♀	(Fe, mmol/l)	7.4 – 10.1	7.5 – 9.9	7.1 – 10.2
PCV ♂	(l/l)	0.41– 0.53	0.41– 0.51	0.40– 0.54
PCV ♀	(l/l)	0.36– 0.49	0.37– 0.47	0.36– 0.47
MCV	(fl)	84 –102	82 –101	76 – 96
MCH	(Fe, fmol)	1.74– 2.14	1.68– 2.11	1.68– 1.99
MCHC	(Fe, mmol/l)	19.6 – 22.1	19.6 – 22.4	18.6 – 21.7

our view that the mean is the same for the two groups. Also the immobilisation period in our hospital is too short to have any influence on the haematopoiesis. It is rather surprising, however, that female blood donors have higher values for haemoglobin, erythrocytes and packed cell volume than female patients. We don't have an explanation for this. The fact that the standard deviation for all parameters is higher for patients than for blood donors can possibly be explained by the fact that the age distribution is much wider for the former group. This fact argues strongly against the use of blood donors for the determination of reference ranges.

The reference values as calculated from a mixed population (14,500 persons, excluding children below the age of 15) are in reasonable agreement with the literature (tab. 5) (8, 9). This population consists of all the patients from whom a sample has been submitted to the laboratory during a 3 month period and comprises about 45% hospitalized-, 45% out patients and 10% blood donors.

An important point indicating that the *Bhattacharya* plot gives correct reference values is the fact that these values are consistent i.e. $\mu_{PCV}/\mu_{RBC} = \mu_{MCV}$, $\mu_{Hb}/\mu_{RBC} = \mu_{MCH}$ and $\mu_{Hb}/\mu_{PCV} = \mu_{MCHC}$. Furthermore the reference range is equally broad for women and men.

For leukocytes it appeared that the frequency distribution could be very well described by the sum of two overlapping *Gaussian* curves. It is not quite clear whether this second curve represents a true subpopulation. The fact, however, that about 30% of all data is part of the second distribution, indicates that they cannot be classified as abnormals. Possibly this group represents persons who suffered from mild infections in the recent past.

Generally it can be concluded that the *Bhattacharya* plot applied to an unselected population is the method of choice for the calculation of reference values for the above mentioned parameters. Smoothing of the frequencies is helpful in deciding which part of the plot is linear when the number of data is small.

References

1. Rutten, W., Scholtis, R., Schmidt, N. & v. Oers, R. (1975), Z. Klin. Chem. Klin. Biochem. *13*, 387–393.
2. Hoeke, J. (1979), Het medisch jaar 1979, 420–429, Bohn, Scheltema & Holkema, Utrecht.
3. Bhattacharya, C. (1967), Biometrics *23*, 115–135.
4. Hoffmann, R. (1963), J. Am. Med. Assoc. *185*, 864–873.
5. White, J. (1978), Clin. Chim. Acta *84*, 353–360.
6. Savitzky, A. & Golay, M. (1964), Anal. Chem. *36*, 1627–1638.
7. Steinier, J., Termonia, Y. & Deltour, J. (1972), Anal. Chem. *44*, 1906–1909.
8. Wintrobe, M. (1974), Clinical Hematology, 7th ed. Lea & Febiger, Philadelphia.
9. Eastham, R. (1974), Clinical Hematology, 4th ed. John Wright & Sons Ltd., Bristol.
10. Diem, K. & Lentner, C. (1969), Wissenschaftliche Tabellen, 7. Auflage, J. R. Geigy AG, Basel.

Ir. A. J. Naus
Dept. of Hematology
and Clinical Chemistry
St. Laurentius Hospital
Mgr. Driessenstraat 6
6043 CV Roermond
The Netherlands